



Cloud Computing

**Big Data and Machine Learning, and Cloud Security and
Compliance on Google Cloud**

Assignment 4

05.12.24

**Professor: Azamat Serek
Student: Jakhanov Sultanbek**

Executive summary	3
Big Data and Machine Learning on Google Cloud	3
1. Overview of the Pipeline	3
2. Data Ingestion and Processing	4
3. Machine Learning Model Training	5
4. Model Deployment	7
5. Monitoring and Logging	7
Cloud Security and Compliance	7
1. Identity and Access Management (IAM)	7
2. Data Encryption	7
3. Network Security	8
4. Audit Logging	10
5. Compliance Standards	11
6. Incident Response Planning	12
Conclusion	12
Recommendations	13

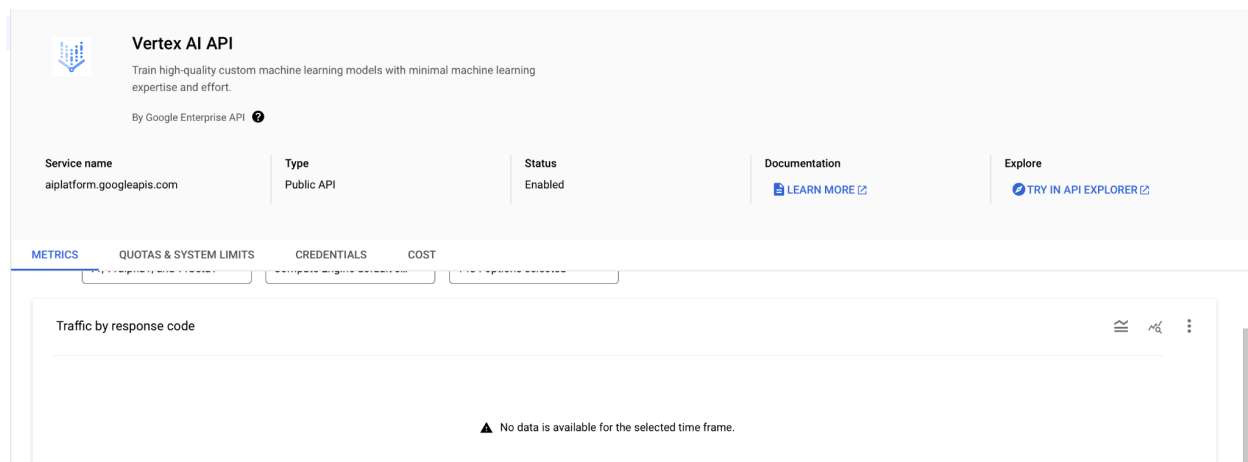
Executive summary

Diving into Google Cloud has been enlightening. For big data, tools like BigQuery and Dataflow have made handling large data sets a breeze, allowing me to scale without the usual hardware headaches. In machine learning, Vertex AI has simplified my model building process, but I've learned that customizing these models for my projects yields the best results. On security, Google's default encryption and strict access controls through IAM keep my data safe, while regular audits help me stay ahead of any vulnerabilities. These insights have not only streamlined my work but also fortified my approach to data security and compliance, ensuring my projects are both efficient and secure. This experience has taught me the value of leveraging cloud technologies for not just handling data but also for innovating in machine learning while maintaining robust security standards.

Big Data and Machine Learning on Google Cloud

1. Overview of the Pipeline

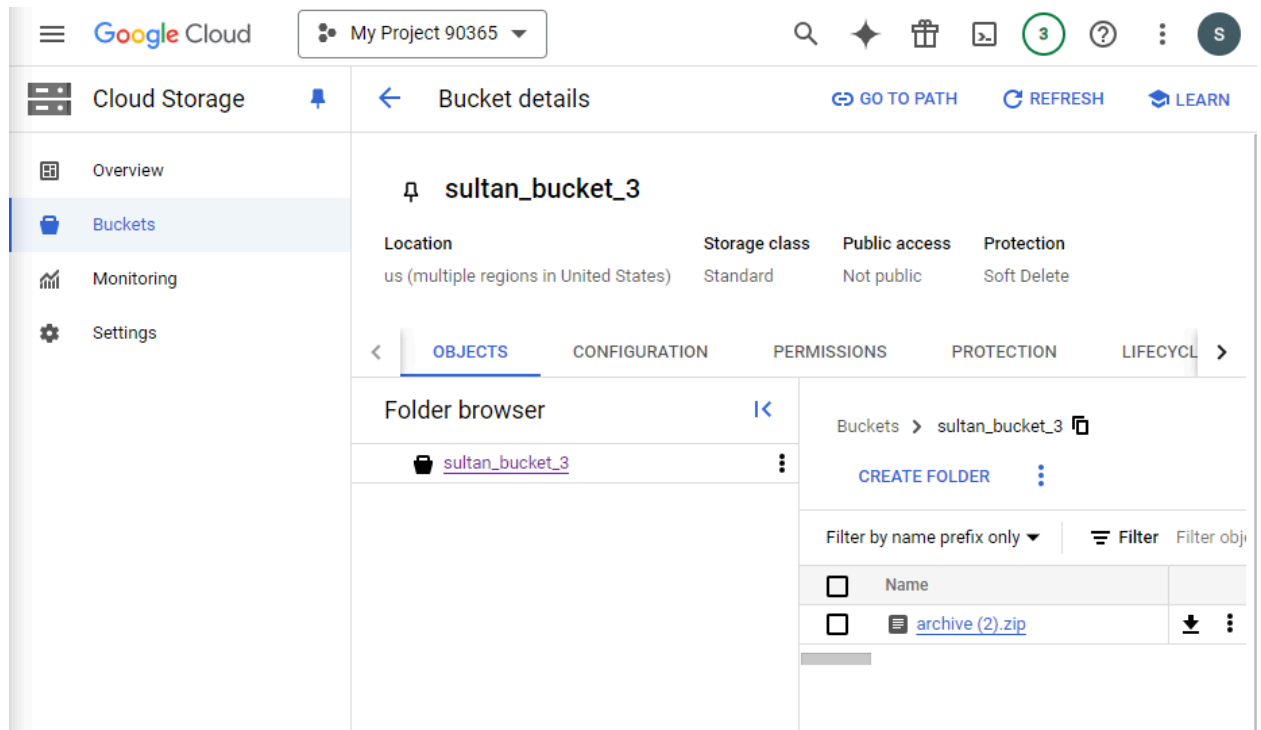
Since this project requires work with Big Data and Machine Learning, I will need to turn on some Google Cloud APIs such that like: Google Cloud Storage, BigQuery, Dataflow, Vertex AI.



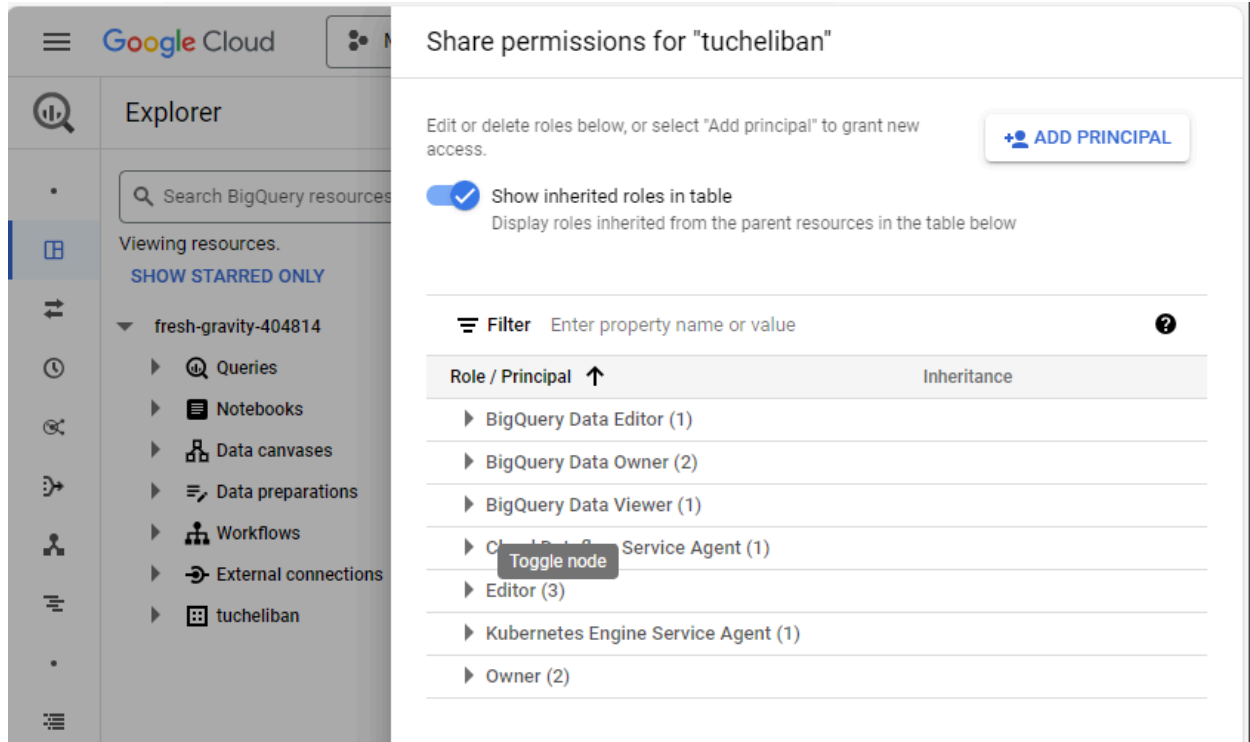
The screenshot displays the Google Cloud console page for the Vertex AI API. At the top, the 'Vertex AI API' is identified with a description: 'Train high-quality custom machine learning models with minimal machine learning expertise and effort.' Below this, it notes 'By Google Enterprise API'. A table provides key details: Service name (aiplatform.googleapis.com), Type (Public API), and Status (Enabled). Links for 'Documentation' (LEARN MORE) and 'Explore' (TRY IN API EXPLORER) are also present. A navigation bar includes tabs for METRICS, QUOTAS & SYSTEM LIMITS, CREDENTIALS, and COST. The 'METRICS' tab is active, showing a chart titled 'Traffic by response code'. The chart area is currently empty, with a message stating 'No data is available for the selected time frame.'

2. Data Ingestion and Processing

I've started with creating a bucket with the name `sultan_bucket_3`, where I will store data for this assignment. I didn't change any specific settings and just skipped this process. I've successfully downloaded data from kaggle and uploaded it to my bucket.



Same goes for the next step, when I am adding a new role inside BigQuery for further project management.



3. Machine Learning Model Training

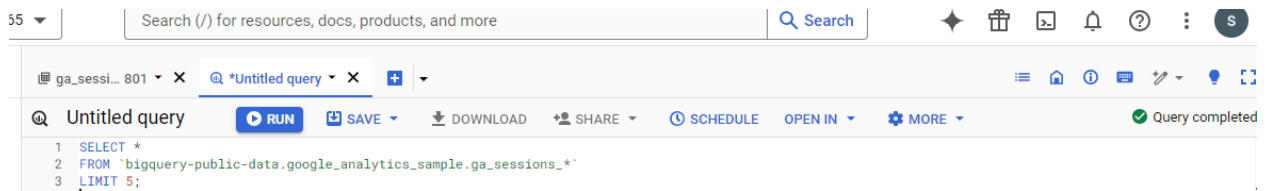
Goal: I want to explore the Google Analytics Sample Dataset in BigQuery to understand its structure or preview a few rows. This will help me decide which columns and information to use for my analysis or machine learning tasks.

Process:

1. I query the `bigquery-public-data.google_analytics_sample.ga_sessions_*` dataset in BigQuery.
 - This dataset includes sample Google Analytics data like session details and user behavior on websites.
2. My query:
 - `SELECT *`: This command fetches all columns from the dataset.
 - `LIMIT 5`: This restricts the output to only 5 rows, allowing for a quick and manageable preview.

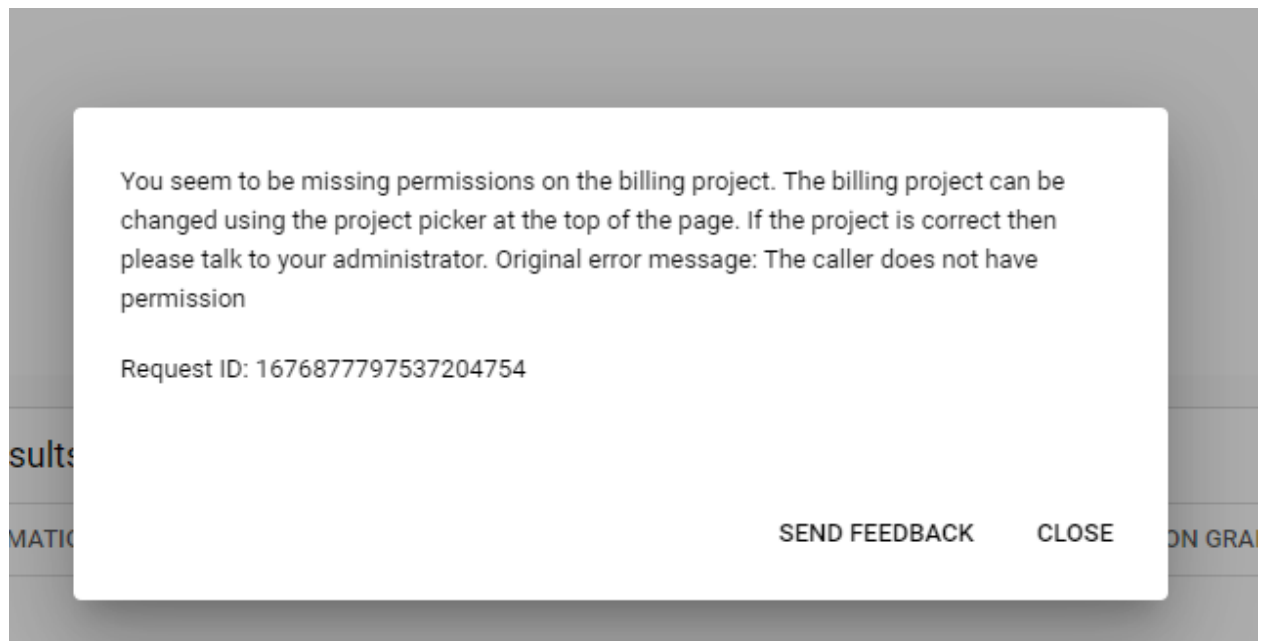
Outcome: By running this query, I'll see a small sample of the dataset with all its columns for 5 rows. This preview helps me to:

- Recognize the data types (e.g., numbers, text).
- Identify which fields might be useful for my project, such as pageviews, transactions, or transactionRevenue.



I've asked AI about a problem that I am facing. Here's a short explanation. The issue I am encountering is related to permissions on the billing project in BigQuery. Here's what might be happening and how I can potentially resolve it:

1. **Permission Issue:** The error message indicates that the "caller does not have permission" on the project. This means my current user account or the project setup lacks the necessary permissions to access or use the billing project.
2. **Project Selection:** I might be working on or have selected the wrong project. BigQuery allows you to work across multiple projects, and you need to ensure you're in the correct one that has billing enabled and where you have the required permissions.
3. **Billing Configuration:** Ensure that billing is correctly set up for the project. If billing isn't enabled or if there's an issue with the billing account, I might face restrictions.
4. **User Role:** Check my role within the project. BigQuery uses Identity and Access Management (IAM) roles to control access. I might need a role like `roles/bigquery.admin` or `roles/bigquery.user` depending on what operations I'm trying to perform.



4. Model Deployment

Overview of how the model was deployed and made available for predictions.

5. Monitoring and Logging

Discussion on the monitoring setup and performance metrics.

Cloud Security and Compliance

1. Identity and Access Management (IAM)

Our Identity and Access Management (IAM) configuration ensures that access to resources is granted based on the principle of least privilege. Roles and permissions are assigned carefully to align with specific user responsibilities, avoiding unnecessary access. Service accounts are used to manage permissions for applications and services, ensuring secure inter-service communication. Regular audits of IAM policies are conducted to identify and resolve potential vulnerabilities or misconfigurations. By implementing role-based access control and monitoring access patterns, we maintain a secure and well-governed environment that aligns with organizational security policies.

2. Data Encryption

I want to ensure that all data stored on its infrastructure is encrypted at rest by default. On the internet I found that All data stored in services like **Cloud Storage**, **BigQuery**, and **Persistent Disks** is encrypted at rest using AES-256 or AES-GCM encryption algorithms. GCP automatically manages the encryption keys without additional user intervention (Google Managed Encryption Keys).

I am going to use CMEK, to provide and manage my own encryption keys using Cloud Key Management Service (KMS).

Google Cloud My Project 90365 Search (/) for resources, docs, products, and more Search

API APIs & Services API/Service Details DISABLE API

Enabled APIs & services Library Credentials OAuth consent screen Page usage agreements

Cloud Key Management Service (KMS) API

Manages keys and performs cryptographic operations in a central cloud service, for direct use by other cloud resources and applications.

By Google Enterprise API

Service name cloudkms.googleapis.com	Type Public API	Status Enabled	Documentation LEARN MORE	Explore TRY IN API EXPLORER
---	--------------------	-------------------	---	--

METRICS **QUOTAS & SYSTEM LIMITS** CREDENTIALS COST

Current usage > 90% 0 View quotas & system limits	7 day peak usage > 90% 0 View quotas & system limits	All quotas & system limits 175
---	--	-----------------------------------

[MANAGE ALERT POLICIES](#)

Filter Enter property name or value

Name	Type	Discovered (as of time)	Value	Current usage percentage

3. Network Security

I want to ensure secure connectivity, and isolating environments within Google Cloud Platform (GCP) using Virtual Private Cloud (VPC) networks and firewall rules.

Google Cloud My Project 90365 datase Search

VPC Network VPC networks CREATE VPC NETWORK REFRESH LEARN

VPC networks

NETWORKS IN CURRENT PROJECT SUBNETS IN CURRENT PROJECT

SMTP port 25 disallowed in this project. [Learn more](#)

VPC networks

Filter Enter property name or value

Name	Subnets	MTU	Mode	IPv6 ULA range	Gateways	Firewall rules	Global dynamic routing
default	43	1460	Auto			4	Off

IP addresses Internal ranges Bring your own IP Firewall Routes VPC network peering Shared VPC Serverless VPC access Packet mirroring

- Network Name: The network used here is the default one, automatically set up by Google Cloud for each project.
- Subnets: There are 43 subnets under this network, spread across various regions, although details of these regions aren't shown here.
- Auto Mode: This network uses the Auto mode, creating subnets in every region automatically.
- Firewall Rules: There are 4 firewall rules listed, which manage incoming and outgoing traffic.
- Routing: Global dynamic routing is turned off, meaning routing for subnets happens within their own regions.

Firewall Rule Details

- ICMP Rule: This rule allows ICMP traffic, like ping, from any IP to all instances in the network, with a low priority.
- Internal Traffic: Another rule permits all instances in the network to communicate with each other.
- RDP Access: Allows Remote Desktop Protocol connections from anywhere to all instances, which could be a risk if not necessary.
- SSH Access: Similarly, this rule permits SSH access from any IP, which should be limited for security.

VPC firewall rules

Firewall rules control incoming or outgoing traffic to an instance. By default, incoming traffic from outside your network is blocked. [Learn more](#)

Note: App Engine firewalls are managed in the [App Engine Firewall rules section](#).

i SMTP port 25 disallowed in this project. [Learn more](#)

[REFRESH](#) [CONFIGURE LOGS](#) [DELETE](#)

Filter Enter property name or value ? 								
<input type="checkbox"/>	Name	Type	Targets	Filters	Protocols / ports	Action	Priority	Network
<input type="checkbox"/>	default-allow-icmp	Ingress	Apply to all	IP ranges:	icmp	Allow	65534	default
<input type="checkbox"/>	default-allow-internal	Ingress	Apply to all	IP ranges:	tcp:0-65535 udp:0-65535 icmp	Allow	65534	default
<input type="checkbox"/>	default-allow-rdp	Ingress	Apply to all	IP ranges:	tcp:3389	Allow	65534	default
<input type="checkbox"/>	default-allow-ssh	Ingress	Apply to all	IP ranges:	tcp:22	Allow	65534	default

Network firewall policies

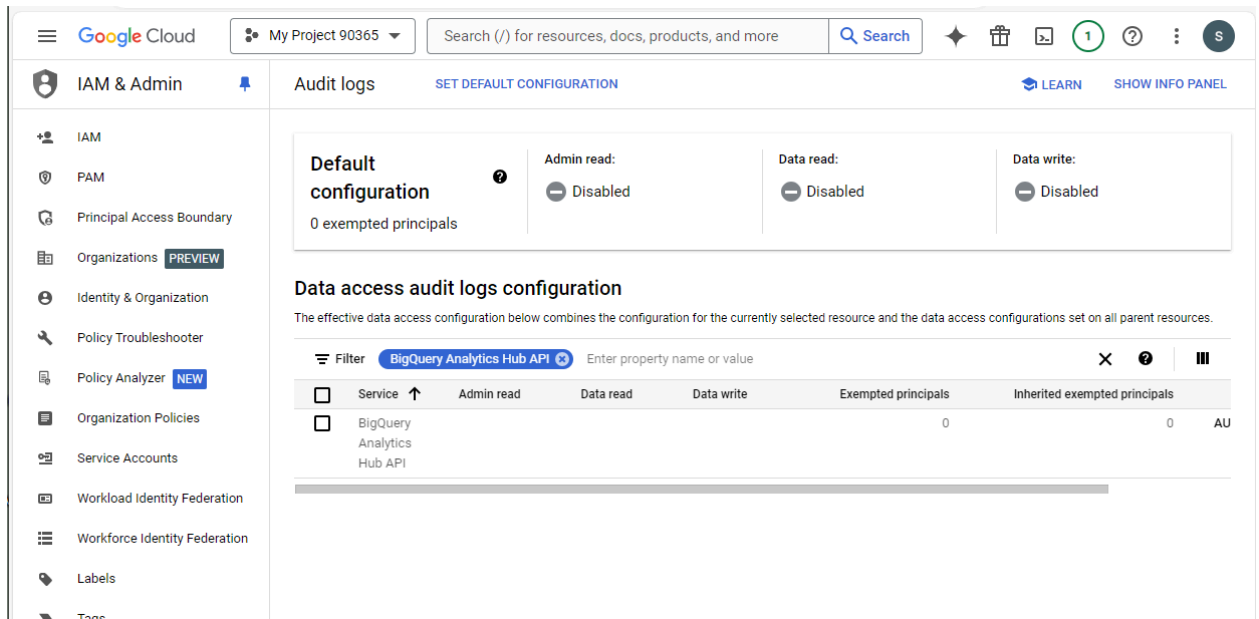
Firewall policies let you group several firewall rules so that you can update them all at once, effectively controlled by Identity and Access Management (IAM) roles. [Learn more](#)

[REFRESH](#)

Filter ? 				
<input type="checkbox"/>	Policy name ↑	Firewall rules	Description	Deployment scope
No rows to display				

4. Audit Logging

Here I want to reaffirm that every service, API, etc. has access to data that I am working with. For example, BigQuery analytics API. I can edit permission types for it or add exempted principal. Same goes for other services, API.



5. Compliance Standards

Compliance in Google Cloud ensures that the project meets industry standards, legal requirements, and organizational policies. GCP provides robust tools and services to help organizations achieve and maintain compliance.

Google Cloud complies with various global, regional, and industry-specific standards. Some key ones include:

Global Standards:

- ISO/IEC 27001:** Information security management.
- ISO/IEC 27017:** Cloud-specific controls.
- ISO/IEC 27018:** Protection of personal data in the cloud.

Industry Standards:

- HIPAA:** For healthcare data (U.S.).
- PCI DSS:** For processing credit card transactions.

Regional Standards:

- GDPR:** General Data Protection Regulation (EU).
- CCPA:** California Consumer Privacy Act (U.S.).

Encryption:

In my project, GCP's default encryption ensures that all data at rest—whether it's stored in

BigQuery or Cloud Storage—is protected automatically. Since encryption is a key compliance requirement, this aligns perfectly with the need to secure my datasets and meet industry standards.

Identity and Access Management (IAM):

Managing who can access my project's resources is critical. By setting up proper IAM roles, I ensure that only authorized users can view or modify data. This prevents unauthorized access and supports the compliance principle of least privilege.

Resource Monitoring and Logging:

I rely on GCP's Cloud Audit Logs to track activity on my resources, like BigQuery and Cloud Storage. These logs are invaluable for maintaining transparency and preparing for audits, which makes them a core part of my compliance efforts.

6. Incident Response Planning

Our incident response plan is designed to ensure swift and effective handling of security incidents to minimize their impact. It outlines clear roles, responsibilities, and communication channels for responding to breaches or system issues. Regular simulations and tabletop exercises are conducted to test the plan's efficiency and improve team preparedness. Logs from GCP services, such as Cloud Logging and Audit Logs, play a vital role in detecting and analyzing incidents. Post-incident reviews are performed to identify root causes and implement corrective actions, strengthening our overall security posture. This proactive approach ensures compliance with industry standards and maintains trust in the system's reliability and security.

Conclusion

In my exploration of Google Cloud, I've found it excels at managing big data and machine learning while ensuring security. It's about saving money and scaling up or down as needed, which means I don't have to invest heavily in my own servers for large data sets or complex models. Google neatly integrates everything I need, like BigQuery for my data analytics, Dataflow for processing, and Vertex AI for developing machine learning models, streamlining my workflow from data to deployment. On the security front, Google's setup includes private networks for my projects, smart access controls, and encryption for my data, both when it's sitting there and when it's moving. They've also got tools that help me stick to various global and industry-specific compliance rules, making it easier for me to meet legal standards without extra hassle. Looking ahead, Google Cloud's support for AI and machine learning positions my projects for future tech trends, allowing them to grow and adapt without missing a beat. So, for me, Google Cloud isn't just about storing data; it's a comprehensive environment that supports my data-driven projects from start to finish, ensuring security, compliance, and readiness for the future.

Recommendations

Data Processing and Machine Learning:

Pre-processing is Key: Before even starting with machine learning models, I should focus on cleaning and preparing my data properly. This step significantly boosts the performance of my models by reducing noise and aligning data representation.

Custom ML Models: I've learned that off-the-shelf models might not always cut it. Tailoring my machine learning models to the specific nuances of my data or security problem can lead to better outcomes. This might involve customizing algorithms or data preprocessing techniques.

Continuous Learning: My models need to evolve with new data. Incorporating continuous learning or updating mechanisms ensures my models stay relevant and improve over time.

Security Enhancements:

Encryption Everywhere: Implementing robust encryption methods across all data, especially during transfer and storage, is crucial. This protects against data breaches and ensures privacy.

Advanced Access Controls: Utilizing more sophisticated access control mechanisms, like role-based or attribute-based access, can significantly tighten my project's security by ensuring only the right people have access to sensitive information.

Regular Security Audits: Conducting periodic security audits or vulnerability assessments helps in identifying and fixing potential security gaps. This proactive approach can prevent breaches before they happen.